Deep learning does not outperform classical machine learning for cell-type annotation

Niklas D. Köhler*[1], Maren Büttner*[1], Niry Andriamanga[1,2], and Fabian J. Theis[1,2,#]

1 Institute of Computational Biology, Helmholtz Center Munich for Environmental Health, Neuherberg, Germany
2 Department of Mathematics, TU Munich, Germany
* these authors contributed equally
[#] e-mail address: fabian.theis@helmholtz-muenchen.de

Abstract

Deep learning has revolutionized image analysis and natural language processing with remarkable accuracies in prediction tasks, such as image labeling and semantic segmentation or named-entity recognition and semantic role labeling. Specifically, the combination of algorithmic and hardware advances with the appearance of large and well-labeled datasets has led up to seminal contributions in these fields.
The emergence of large amounts of data from single-cell RNA-seq and the recent global effort to chart all cell types in the Human Cell Atlas has attracted an interest in deep-learning applications. However, all current approaches are unsupervised, *i.e.*, learning of latent spaces without using any cell labels, even though supervised learning approaches are often more powerful in feature learning and the most popular approach in the current AI revolution by far. Here, we ask why this is the case. In particular we ask whether supervised deep learning can be used for cell annotation, *i.e.* to predict cell-type labels from single-cell gene expression profiles. After evaluating 10 classification methods across 14 datasets, we notably find that deep learning does not outperform classical machine-learning methods in the task. Thus, cell-type prediction based on gene-signature derived cell-type labels is potentially too simplistic a task for complex non-linear methods, which demands better labels of functional single-cell readouts.

Main

Single-cell genomics is a success story in global scientific collaborative activities and research, combining advances in genomics, miniaturization, and data science. The widespread availability of transcriptome readouts from various cells, tissues, and species nowadays allows the integration and comparison of such data using machine-learning methods[1–5]. Successful data integration in turn facilitates reliable annotation of cell types at new datasets levels of quality as compared with those of human experts. Annotating cell types is a classical supervised learning problem, which can be addressed by using numerous classification

models, such as logistic regression, support vector machines, and random forests. However, with increasing data volumes, we can soon approach data volume thresholds where deep learning models outperform classical machine-learning models in cell-type annotation tasks. Here we explore whether we have reached this threshold by comparing classical machine-learning approaches with state-of-the-art deep-learning architectures in a cell-type annotation task. In our comparison study of classifier types, we present a strategy for deriving optimal parameters for a given model to predict cell-type annotation, *i.e.*, what level of accuracy can we expect in the best case scenario from a given model trained using a specific dataset. We applied this testing framework to three different biological systems with increasing sample sizes: mouse embryonic development[6–10] (600 cells), human pancreas[11–14] (approx. 20,000 cells), and mouse brain[15–18] (approx. 1,000,000 cells), each with varied complexity.

A unified strategy for cell type annotation

We consider the following general problem setup: Given the access to a set of *n-1* reference studies with labeled cell types, our objective is to assign unknown cells to cell types automatically in a new target study of interest. Notably, recent studies for cell type annotation tasks did not split by dataset for evaluating the prediction accuracy, therefore limiting the informative power [19–21]. Predicting a previously unseen dataset is a classical supervised learning problem with the added complexity that, depending on the study, we may expect novel cell types (labels). We address the issue in a manner similar to scmap[1] as follows. First, we define a confidence threshold in our prediction (*thresh = 0.5*). Please note that every cell-type class prediction has a specific confidence level. Subsequently, we assign cells below the threshold with the tag “unknown cell type”, which indicates that the cell has minimal similarity to all the annotated cell types in the reference studies, and is therefore, potentially a novel cell type.

We analyzed five machine-learning models of increasing complexity for our comparison: k-NN classification (knn), logistic regression (lgr) (with both Lasso and ridge penalty), Support Vector Machines (svm), Gradient Boosted Trees (xgb), and Multilayer Perceptrons (mlp) (see **Supplementary Methods** for details). To ensure a meaningful and unbiased comparison, we perform a five-fold cross-validated grid search over a wide range of potential parameters for all classifiers (see **Supplementary Fig. 1**). The optimal parameter set for each classifier is selected based on the highest mean accuracy score averaged over classes and over the respective five-fold cross-validation test set predictions, therefore accounting for class imbalance. Notably, using class weighting instead did not improve the average accuracy nor the overall accuracy (see **Supplementary Table 1**). We perform the grid search approach in a leave-one-study-out fashion for all three biological systems, resulting in 14 five-fold cross-validated grid searches for five different classifier types. The classifiers with the final parameter sets are then used to predict the unknown cell types in the target study.

We further included five single-cell specific cell type annotation models: scmap[1], SingleR[22], Garnett[5], scMatch[23] and cellFishing.jl[24]. While scmap, SingleR and cellFishing.jl can be trained on *n-1* datasets and applied to the *n*th dataset, scMatch uses a FANTOM5 database derived reference and Garnett trains a classifier on marker genes.

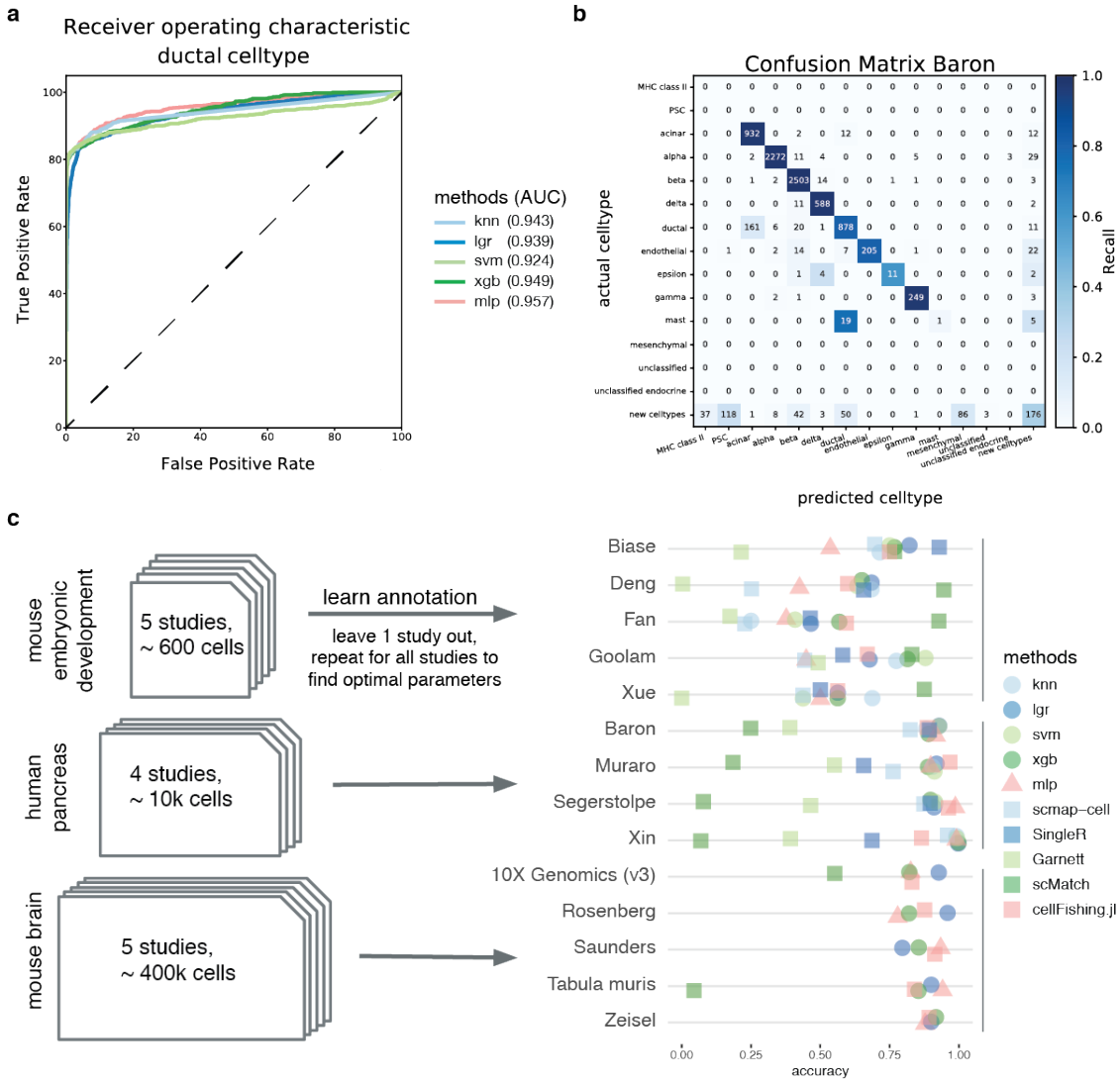

Figure 1: All supervised learning models for cell-type annotation have similar performance as measured by accuracy. (a) *Receiver operating characteristics for the ductal cell type (Baron data as holdout study) in all models, which return prediction probability for all classes. The area under curve (AUC) is given in parentheses in the legend. We observe little difference across classifier types.* ***(b)*** *Confusion matrix for the recall of the MLP in pancreas data with Baron12 data as holdout study.* ***(c)*** *Accuracy scores for knn, lgr, svm, xgb (circles) and mlp (red triangles) in each study using the optimal parameter set and single-cell specific models (squares, scmap-cell, SingleR, Garnett, scMatch and cellFishing.jl). In the case of mouse brain data, only included lgr, xgb, mlp, cellFishing.jl and scMatch (2 data sets) due to infeasible runtime or memory usage.*

Most machine learning methods are well-suited for cell type annotation given optimal hyper-parameter values

We evaluated cell-type annotation performance of models based on the accuracy of predicting cell types in a holdout study. In the case of scMatch, we considered a cell as correctly classified, if it was assigned to the correct tissue (e.g. embryonic tissue in the case of embryonic development and pancreas for the human pancreas scenario). For the machine-learning models, we particularly measured the accuracy based on receiver operator characteristics per cell type to determine the best performing model across all classifier types (see **Fig. 1a**). Subsequently, we determined the recall of the best performing model and present it as a confusion matrix of actual and predicted cell types in the target study (see **Fig. 1b** and **Supplementary Figure 2**). The final evaluation score is the accuracy of the top

performing classifier for predicting all cell-type classes calculated on the target study (see **Fig. 1c**).

For the small-scale embryonic development datasets, we observed considerable differences in performance across models in the same dataset, which could be explained by the small sample size of the training data. For example, the dataset from Deng et al.[8] comprises 314 cells, and the training dataset comprises 284 cells. However, the overall accuracy of the prediction across datasets was not correlated with the sample size ratio of the training and test dataset. Instead, the differences in performance across models on the same test dataset decreased with increasing sample size in the training dataset.

When we used tissue studies of the human pancreas[11–14] and the mouse brain[15–18], all machine-learning models and *cellFishing.jl* performed equally well on the prediction task. Smith et al.[25] reported a similar result in bulk RNA-seq data. In contrast, the single-cell specific classifiers *scMatch* and *Garnett* showed lower accuracy on the same datasets. We achieved a high prediction accuracy and recall in human pancreas studies. We have illustrated the cell-type prediction for the Baron[12] dataset with MLP (see **Fig. 1b**). Notably, MLP could not predict several rare cell types at all (e.g. mast cells), while it predicted abundant cell types accurately. Notably the reference dataset was more complex and contained more cell types compared to the predicted dataset. The prediction favored rare cell types, such as mesenchymal cells and PSCs, which the authors of the original publication did not observe. Whether the prediction reflects the ground truth better than the human expert labels, is debatable. In general, reproducibility across human experts is limited [26] and algorithmic discovery of rare cell types remains challenging [27].

In case of the large mouse brain datasets, we included lgr, xgb, mlp, *cellFishing.jl* and *scMatch* (2 data sets) due to limitations in runtime and memory usage of all other models. The accuracy scores for each classifier varied minimally for every holdout study and were generally comparable across all holdout studies. We even observed a good performance on the 10X Genomics prenatal mouse brain dataset (E18 stage) even though we expected differences caused by maturation from the training data, which were obtained from postnatal mice. The developing brain undergoes considerable maturation processes before and after birth, which are reflected in the transcriptomes of the respective cells. Surprisingly, the lower transcriptional similarity between test data and training data did not result in lower prediction accuracy scores.

Overall, we found that the deep learning models did not outperform classical machine-learning models, such as logistic regression and svm (pairwise wilcoxon rank sum test, see **Supplementary Tables 2-4**), in the prediction/annotation of cell types in new data sets. Instead, factors, such as similarities between training and test data, data complexity in terms of cell types, and mere sample size affect the accuracy of annotation. Nevertheless, model tuning through grid search considerably improved accuracy. Almost all tested classifiers significantly outperformed *scmap-cell*[1] and *Garnett*[5] based on accuracy scores (pairwise wilcoxon rank sum test, see **Supplementary Tables 2-4**, see **Fig. 1c**). Interestingly, cellFishing.jl[24] performed equally well as the general classifiers, significantly better than *Garnett* (see **Supplementary Table 2**), and was the only single-cell specific tool that scaled to all datasets in this study.

Discussion and conclusion

We tested several supervised classifiers in small, medium, and large-scale datasets. We compared optimally tuned classification models with various degrees of complexity using a

global hyperparameter search to obtain optimal parameter sets. It is not surprising that the performance of all the trained models improved significantly compared to the model *scmap-cell,* which uses a fixed set of parameters. Clearly, cell-type annotation accuracy benefits from the screening for optimal parameters independent of the model. In addition, the accuracy of the various classifiers is largely similar for each holdout study across numerous data sets, where the differences in the large datasets are lower compared to the differences in small datasets. Interestingly, the feature-based classifier Garnett and the database approach scMatch were significantly outperformed by supervised classification with cell type labels.

Why does deep learning not outperform classical machine-learning methods, in particular for the large and complex droplet-based data sets with more than half a million cells? Firstly, our input data is tabular rather than sequential. Additional information on the role of genes, as encoded in gene regulatory networks, could be leveraged e.g. in graph convolutional networks. Secondly, we have to consider when a learning task is 'complex' enough for powerful representation learning algorithms like deep learning to outperform simpler methods like kNN, logistic regression and SVMs. The difference between a simple and a complex learning task lies in the complexity of the transformations of the raw data necessary to identify a representation, which linearly separates the target labels of interest. Therefore, a learning task can be considered 'complex' when its label-separating representation has very little resemblance with the similarity structure present in the raw data. For example, learning on RGB image datasets is simple when we want to predict the amount of red vs green pixels, but a learning task on the same image dataset becomes incredibly complex when we aim to predict the content on an image. Deep learning is a very powerful representation learning technique, which excels exactly in constructing such representations: Given enough data, deep learning based methods are able to discover features, which span vastly different data graphs compared to raw data. We illustrated this well-known feature on the often-studied cifar10 dataset (60,000 images in 10 mutually exclusive classes; see **Supplementary Fig. 3**). Here, separating the classes based on the raw input is unsuccessful (**Supplementary Fig. 3a**) while the neural network recovers a clean separation between the classes (**Supplementary Fig. 3b**). In the case of image data, a deep learning model creates a nonlinear connection to a simpler latent space, where decision boundaries accurately discriminate between the complex labels. In contrast to this image-content prediction example, the annotation of cell types in single cell data relies on the cluster analysis of a feature reduced space (e.g. through selecting highly variable genes and performing PCA as a pre-processing step). Here, the cell type label is a non-complex label in the sense that the activity of a few key marker genes provides an often-sufficient grouping of similar cells, *i.e.* the distance on the reduced feature space induces an already biologically interpretable structure of the data.

Hence, to be able to use deep-learning-based representations to discover non-trivial and generalizable feature spaces, particularly interesting for subsequent transfer learning[28], more complex tasks should be addressed. Such complex functional relationships are encountered between the cellular state (cell type or cluster identity) and functional assays, responses to drug treatment or gene knockouts[29,30]. Furthermore, complex labels such as cell fates require creative experimental methods for tracing over several generations in both cell lines and complete animals, such as traceable genetic manipulation[31,32]. In tissue samples, the transcriptome of cells depends on its direct neighbourhood[33,34], *i.e.* when we combine spatial information with RNA-seq, we can ask to predict a cell state together as a function of its neighborhood as in local frames in image data. Furthermore, paired multi-omics data such as

CITE-Seq[35] allow the prediction of protein abundances, where deep learning models are already applied[36]. Altogether, when labels become more complex, we expect the benefits of deep learning based classifiers to increase.

While waiting for the emergence of such labels, the scientific community has developed a series of unsupervised deep-learning approaches, including batch effect correction[37], large scale data integration[38], single-cell perturbation prediction[39], and dropout effects denoising[40]. However, the prediction of cell types is cell based, not cluster based. Therefore, a better clustering does not address the lack of complexity in the cell type annotation, but it reduces noise, increases accuracy and reduces occasional inconsistencies.
In general, the application of neural networks in supervised cell-type annotation provides practical advantages in terms of flexible applicability, highly optimized training libraries, and access to learned feature representation in latent spaces for downstream analysis and visualization[28].
In the present study, we were not able to observe any significant improvement in classification accuracy based on neural networks, in contrast to other fields, such as computer vision[41] and natural language processing[42–44]. Therefore, if we are to achieve the “ImageNet moment” in single-cell genomics, where end-to-end deep-learning models heavily outperform classical machine learning methods in the limit of large training datasets, we will need to work on better annotation and data with more covariates such as replicates, samples or patients in the future.

References

1. Kiselev, V. Y., Yiu, A. & Hemberg, M. scmap: projection of single-cell RNA-seq data across data sets. *Nat. Methods* **15**, 359–362 (2018).
2. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36**, 421–427 (2018).
3. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
4. Stuart, T. *et al.* Comprehensive integration of single cell data. *Cell* **177**, 1888–1902.e21 (2019).
5. Pliner, H. A., Shendure, J. & Trapnell, C. Supervised classification enables rapid annotation of cell atlases. *Nature Methods* **16,** 983–986 (2019).
6. Biase, F. H., Cao, X. & Zhong, S. Cell fate inclination within 2-cell and 4-cell mouse embryos revealed by single-cell RNA sequencing. *Genome Res.* **24**, 1787–1796 (2014).

7. Goolam, M. *et al.* Heterogeneity in Oct4 and Sox2 Targets Biases Cell Fate in 4-Cell Mouse Embryos. *Cell* **165**, 61–74 (2016).
8. Deng, Q., Ramsköld, D., Reinius, B. & Sandberg, R. Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science* **343**, 193–196 (2014).
9. Xue, Z. *et al.* Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature* **500**, 593–597 (2013).
10. Fan, X. *et al.* Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. *Genome Biol.* **16**, 148 (2015).
11. Segerstolpe, Å. *et al.* Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.* **24**, 593–607 (2016).
12. Baron, M. *et al.* A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell Syst* **3**, 346–360.e4 (2016).
13. Xin, Y. *et al.* RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes. *Cell Metab.* **24**, 608–615 (2016).
14. Muraro, M. J. *et al.* A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Syst* **3**, 385–394.e3 (2016).
15. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
16. Zeisel, A. *et al.* Molecular Architecture of the Mouse Nervous System. *Cell* **174**, 999–1014.e22 (2018).
17. Tabula Muris Consortium *et al.* Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
18. Saunders, A. *et al.* Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. *Cell* **174**, 1015–1030.e16 (2018).
19. Abdelaal, T. *et al.* A comparison of automatic cell identification methods for single-cell RNA sequencing data. *Genome Biol.* **20**, 194 (2019).
20. Shao, X. *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-

Cell RNA Sequencing Data. *iScience* **23**, 100882 (2020).

21. Li, C. *et al.* SciBet as a portable and fast single cell type identifier. *Nat. Commun.* **11**, 1818 (2020).
22. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
23. Hou, R., Denisenko, E. & Forrest, A. R. R. scMatch: a single-cell gene expression profile annotation tool using reference datasets. *Bioinformatics* **35**, 4688–4695 (2019).
24. Sato, K., Tsuyuzaki, K., Shimizu, K. & Nikaido, I. CellFishing.jl: an ultrafast and scalable cell search method for single-cell RNA sequencing. *Genome Biol.* **20**, 31 (2019).
25. Smith, A. M. *et al.* Deep learning of representations for transcriptomics-based phenotype prediction. *bioRxiv* (2019) doi:10.1101/574723.
26. Valizadegan, H., Nguyen, Q. & Hauskrecht, M. Learning classification models from multiple experts. *J. Biomed. Inform.* **46**, 1125–1135 (2013).
27. Jindal, A., Gupta, P., Jayadeva & Sengupta, D. Discovery of rare cells from voluminous single cell expression data. *Nat. Commun.* **9**, 4719 (2018).
28. Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nat. Rev. Genet.* (2019) doi:10.1038/s41576-019-0122-6.
29. Dixit, A. *et al.* Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
30. Datlinger, P. *et al.* Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
31. Chan, M. M. *et al.* Molecular recording of mammalian embryogenesis. *Nature* (2019) doi:10.1038/s41586-019-1184-5.
32. Wagner, D. E. *et al.* Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science* **360**, 981–987 (2018).
33. Satija, R., Farrell, J. A., Gennert, D. & Schier, A. F. Spatial reconstruction of single-cell gene expression data. *Nature* (2015).

34. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4231.
35. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
36. Zhou, Z., Ye, C., Wang, J. & Zhang, N. R. Surface protein imputation from single cell transcriptomes by deep neural networks. *bioRxiv* 671180 (2019) doi:10.1101/671180.
37. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
38. Luecken *et al.* Benchmarking atlas-level data integration in single-cell genomics. *biorxiv* (2020) doi:10.1101/2020.05.22.111161.
39. Lotfollahi, M., Alexander Wolf, F. & Theis, F. J. scGen predicts single-cell perturbation responses. *Nature Methods* **16**, 715–721 (2019).
40. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nat. Commun.* **10**, 390 (2019).
41. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).
42. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training. (2018).
43. Peters, M. E. *et al.* Deep contextualized word representations. *arXiv [cs.CL]* (2018).
44. Collobert, R. *et al.* Natural language processing (almost) from scratch. *J. Mach. Learn. Res.* **12**, 2493–2537 (2011).